Tradict enables high fidelity reconstruction of the eukaryotic transcriptome from 100 marker genes

Surojit Biswas[1,*], Konstantin Kerner[2], Paulo José Pereira Lima Teixeira[3,4], Jeffery L. Dangl[3-7,†], Vladimir Jojic[8,†], Philip A. Wigge[9,†,*]

[1] Department of Biomedical Informatics, Harvard Medical School, Boston, MA 02115, USA
[2] Botanical Institute, Biocenter, University of Cologne, D-50674 Cologne, Germany.
[3] Howard Hughes Medical Institute and [4] Department of Biology, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA.
[5] Carolina Center for Genome Science, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA
[6] Department of Microbiology and Immunology, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA
[7] Curriculum in Genetics and Molecular Biology, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA
[8] Department of Computer Science, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA
[9] Sainsbury Laboratory, University of Cambridge, Cambridge CB2 1LR, UK

[†] Contributed equally
* Correspondence: surojitbiswas@g.harvard.edu, Philip.Wigge@slcu.cam.ac.uk

Abstract

Transcript levels are a critical determinant of the proteome and hence cellular function. Because the transcriptome is an outcome of the interactions between genes and their products, we reasoned it might be accurately represented by a subset of transcript abundances. We develop a method, Tradict (transcriptome predict), capable of learning and using the expression measurements of a small subset of 100 marker genes to reconstruct entire transcriptomes. By analyzing over 23,000 publicly available RNA-Seq datasets, we show that Tradict is robust to noise and accurate, especially for predicting the expression of a comprehensive, but interpretable list of transcriptional programs that represent the major biological processes and cellular pathways. Coupled with targeted RNA sequencing, Tradict may therefore enable simultaneous transcriptome-wide screening and mechanistic investigation at large scales. Thus, whether for performing forward genetic, chemogenomic, or agricultural screens or for profiling single-cells, Tradict promises to help accelerate genetic dissection and drug discovery.

Introduction

As the critical determinant of the proteome and therefore cellular status, the transcriptome represents a key node of regulation for all life[1]. Transcriptional control is managed by a finely tuned network of transcription factors that integrate environmental and developmental cues in order to actuate the appropriate responses in gene expression[2–4]. Importantly, the transcriptomic state space is constrained. Pareto optimality suggests that no gene expression profile or phenotype can be optimal for all tasks, and consequently, that some expression profiles or phenotypes must come at the expense of others[5,6]. Furthermore, across all major studied kingdoms of life, cellular networks demonstrate remarkably conserved scale-free properties that are topologically characterized by a small minority of highly connected regulatory nodes that link the remaining majority of sparsely connected nodes to the network[7–9]. These theories suggest that the effective dimension of the transcriptome should be far less than the total number of genes it contains. If true to a large enough extent, it may be possible to faithfully compress and prospectively reconstruct the entire transcriptomes using only a small, carefully chosen subset of it.

Indeed, previous studies have exploited this reduced dimensionality to perform gene expression imputation for microarray data for missing or corrupted values[10–12]. Others have extended these intuitions to predict expression from probe sets containing a few hundred genes[13,14]. However, prediction accuracies have been modest and usually limited to 4,000 target probes/genes. Recently, several studies examined the transcriptomic information recoverable by shallow sequencing especially as it applies to single-cell experiments[15–18]. Jaitin *et al.* (2014) and Pollen *et al.* (2014) demonstrated that only tens of thousands of reads are required per cell to learn and classify cell types

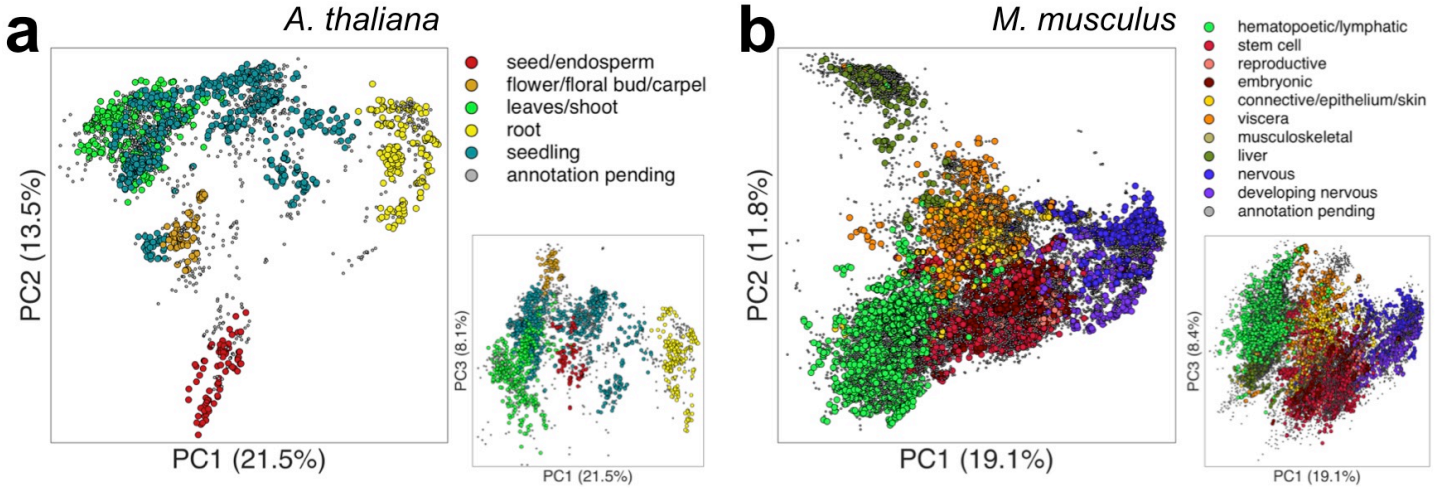

Figure 1. The primary drivers of variation in our training transcriptome collection are developmental stage and tissue. a) *A. thaliana*, b) *M. musculus*. Also shown are plots of PC3 vs. PC1 to provide another perspective.

ab initio[16,18]. Heimberg *et al.* (2016) extended these findings and demonstrated that the major principal components of a typically sequenced mouse bulk or single-cell expression dataset may be estimated with little error at even 1% of the depth[15]. Though these approaches advance the notion of strategic transcriptome undersampling, they only recover broad transcriptional states and are restricted to measuring only the most abundant genes. During sample preparation -- arguably the most expensive cost of a multiplexed sequencing experiment -- shallow sequencing-based approaches still utilize protocols meant for sampling the entire transcriptome and thus consume more resources than necessary. Furthermore, given that the expression of even the most abundant genes is highly skewed, sequencing effort is wastefully distributed compared to an approach that chooses which genes to measure more wisely. Finally, it is still not clear from sample sizes and biological contexts previously studied whether the low dimensionality of the transcriptome may be leveraged unconditionally (or nearly so) across organism and application.

In this work, we introduce Tradict (transcriptome predict), a novel, robust-to-noise, and probabilistically sound algorithm for inferring the transcriptome using only the expression measurements of a single, context-independent, machine-learned subset of 100 marker genes. Using a transcriptionally representative sampling of over 23,000 publicly available, transcriptome-wide RNA-Seq datasets for *Arabidopsis thaliana* and *Mus musculus*, we train Tradict to prospectively reconstruct gene expression, and to predict, with a striking degree of accuracy, the expression of a comprehensive, but quickly interpretable collection of transcriptional programs that represent the major biological processes and pathways of the cell. Our work demonstrates the development and large-scale application of a multivariate count/non-negative data model, and highlights the power of directly modeling the expression of transcriptional programs in a supervised manner. We believe that Tradict, coupled with targeted RNA sequencing[19–22], can rapidly illuminate biological mechanism and improve the time and cost of performing large forward genetic, breeding, or chemogenomic screens, accurately profiling single-cells, and performing gene expression based clinical diagnostics.

Results

Assembly of a comprehensive training collection of transcriptomes

We attempted to download all available Illumina sequenced publicly deposited RNA-Seq samples (transcriptomes) on NCBI's Sequence Read Archive (SRA). Among samples with at least 4 million reads, we successfully downloaded and quantified the raw sequence data of 3,621 and 27,450 transcriptomes for *A. thaliana* and *M. musculus,* respectively. After stringent quality filtering, we retained 2,597 (71.7%) and 20,847 (76.0%) transcriptomes comprising 225 and 732 unique SRA submissions for *A. thaliana* and *M. musculus*, respectively. An SRA 'submission' consists of multiple, experimentally linked samples submitted concurrently by an individual or lab. We defined 21,277 (*A. thaliana*) and 21,176 (*M. musculus)* measurable genes with reproducibly detectable expression given our tolerated minimum sequencing depth and mapping rates (see Supplemental Information "Materials and Methods" for further information regarding data acquisition, transcript quantification, quality filtering, and

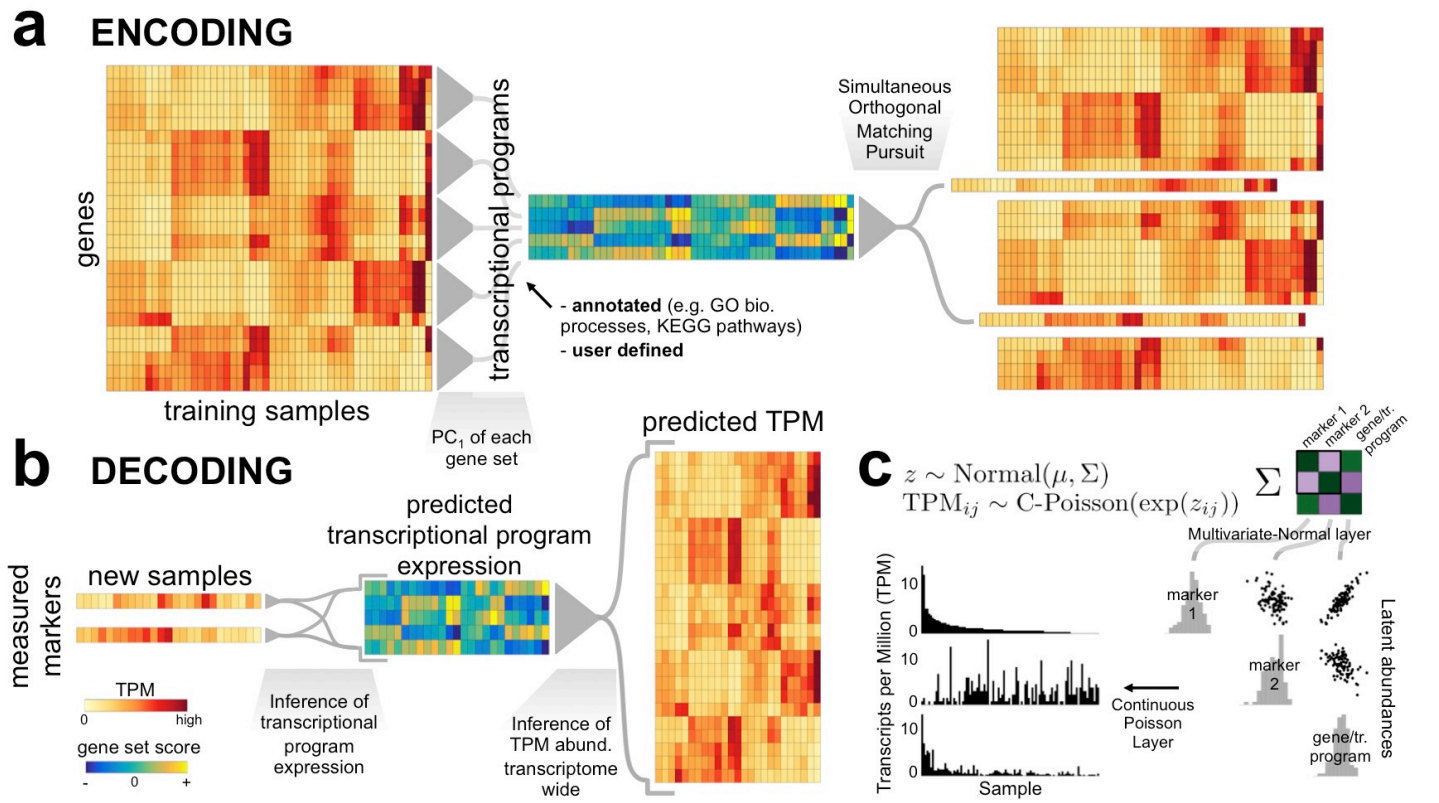

Figure 2. Tradict's algorithmic workflow. a) During encoding the transcriptome is first quantitatively summarized in terms of a collection of a few hundred, biologically comprehensive transcriptional programs. These are then decomposed into a subset of marker genes using an adaptation of the Simultaneous Orthogonal Matching Pursuit algorithm, and a Multivariate Normal Continuous-Poisson hierarchical model is used as a predictive model to capture covariance relationships between markers, transcriptional programs, and all genes. b) During decoding, Tradict predicts the expression of transcriptional programs and all genes in the transcriptome using only expression measurements of the marker genes. c) The Multivariate Normal Continuous-Poisson hierarchy enables Tradict to efficiently model statistical coupling between the non-negative expression measurements typical of sequencing. This is done by assuming that associated with each observed, noisy TPM measurement, there is an unmeasured, denoised latent abundance the logarithm of which comes from a Multivariate Normal distribution over all genes and transcriptional programs.

expression filtering). We hereafter refer to the collection of quality and expression filtered transcriptomes as our *training transcriptome collection*.

In order to assess the quality and comprehensiveness of our training collection, we performed a deep characterization of the expression spaced spanned by these transcriptomes. We found that the transcriptome of both organisms was highly compressible and that the primary drivers of variation were tissue and developmental stage (Figure 1a-b, Figure S1), with many significant, biologically realistic trends within each cluster (Supplemental Note 1). We additionally examined the distribution of submissions across the expression space, compared inter-submission variability within and between tissues, inspected expression correlations among genes with well-established regulatory relationships, and assessed the evolution of the expression space across time. These investigations revealed our training collection is of high and reproducible technical quality, reflective of known biology, and stable (Supplemental Note 1, Figures S2-S4). Given additionally the diversity of tissues, genetic perturbations, and environmental stimuli represented in the SRA, these results, taken together, suggest that our training collection is an accurate and representative sampling of the transcriptomic state space that is of experimental interest for both organisms.

Tradict - algorithm overview

Given a training sample of transcriptomes, Tradict encodes the transcriptome into a single subset of globally representative *marker* genes and learns their predictive relationship to the expression of a comprehensive collection of transcriptional programs (e.g. pathways, biological processes) and to the rest of the transcriptome. Tradict's key innovation lies in using a Multivariate Normal Continuous-Poisson (MVN-CP) hierarchical model to model marker latent abundances -- rather than their measured, noisy abundances -- jointly with the expression of transcriptional programs, and ultimately, the latent abundances of the remaining non-marker genes in the transcriptome. In so doing,

Tradict is able to 1) efficiently capture covariance structure within the non-negative, right-skewed output typical of sequencing experiments, and 2) perform robust inference of gene set and non-marker expression even in the presence of significant noise.

Figure 2 illustrates Tradict's general workflow. Estimates of expression are noisy, especially for low to moderately expressed genes. Given samples are often explored unevenly and that the *a priori* abundance of each gene differs, the level of noise in a gene's measured expression for a given sample varies, but it can be estimated. Therefore, during training, Tradict first learns the log-latent, denoised abundances for each gene in every sample in the training collection using the lag transformation[23]. It then collapses this latent transcriptome into a collection of predefined, globally comprehensive collection of *transcriptional programs* that represent the major processes and pathways of the cell related growth, development, and response to the environment (Supplemental Tables 3-4). In this work, we focus on creating a Gene Ontology derived panel of transcriptional programs, in which the first principal component of all genes contained within an appropriately sized and representative GO term is used to define the similarly named transcriptional program[24,25]. The expression values of these programs are then encoded using an adapted version of the Simultaneous Orthogonal Matching Pursuit into a small subset of marker genes selected from the transcriptome[26,27]. Tradict finally stores the mean and covariance relationships between the log-latent expression of the selected markers, the transcriptional programs, and the log-latent expression of the remaining non-marker genes at the Multivariate Normal layer of the underlying MVN-CP hierarchical model for use in future decoding (Figure 2a).

Prospectively, only the expression of these marker genes needs to be measured and the expression of transcriptional programs and/or the rest of the transcriptome can be inferred as needed. During this process of decoding, Tradict first utilizes an iterative Bayesian updating algorithm to learn the log-latent abundances associated with each measured maker for every sample. Though a simply a consequence of proper inference of our model, this denoising step adds considerable robustness to Tradict's predictions. Tradict then uses the covariance relationships learned during training to formulate a prediction for the expression of transcriptional programs and the most likely expression values for all remaining non-marker genes (Figure 2b).

Tradict prospectively predicts unseen transcriptomes and transcriptional programs with robust accuracy

An SRA submission consists of multiple, experimentally linked samples submitted concurrently by an individual or lab. Consequently, high *intra*-submission prospective prediction accuracy is most indicative of a method's performance.

To more completely understand Tradict's prospective predictive performance, we performed 20-fold cross validation on the training transcriptome collections for both *A. thaliana* and *M. musculus* and evaluated Pearson correlation coefficients (PCC) between predicted and actual expression for each fold when the remaining 95% of folds were used for training. To make this experiment as reflective of reality as possible, folds were divided by submission so that samples from the same set of experiments would not appear both in training and test sets. Because submissions to the SRA span a broad array of biological contexts, the total biological signal contained in the test set exceeds that of what would be expected for typical application, which in turn would lead to overly optimistic estimates of prediction accuracy. To therefore evaluate *intra-submission* accuracy, PCC calculations were performed on 'submission-adjusted' expression values in which each submission's mean expression was subtracted from the expression values of all associated samples.

Figures 3a and 3c illustrate that the reconstruction performance for transcriptional programs in both organisms is strikingly accurate across all collected submissions. Quantitatively speaking, the average intra-submission PCCs for transcriptional programs are 0.94 and 0.93 for *A. thaliana* and *M. musculus*, respectively. This is despite lower, but still accurate prediction performance on gene expression (Figures 3b and 3d). Intuitively, this because transcriptional programs are measured as linear combinations of the log-latent TPMs of the genes that comprise them, effectively smoothing over the orthogonal noise present in each gene's expression prediction.

Gene expression prediction error was negatively correlated with mean expression (ρ = -0.496 *A. thaliana*, ρ = -0.607 *M. musculus*; Spearman correlation) consistent with the findings of Donner *et al.* 2012. Similarly, for transcriptional programs, prediction error was negatively correlated with the mean expression of genes (ρ = -0.325 *A. thaliana*, ρ = -0.577 *M. musculus*; Spearman correlation) and the number of genes (ρ = -0.545 *A. thaliana*, ρ = -0.5826 *M. musculus*; Spearman correlation) contained within each program

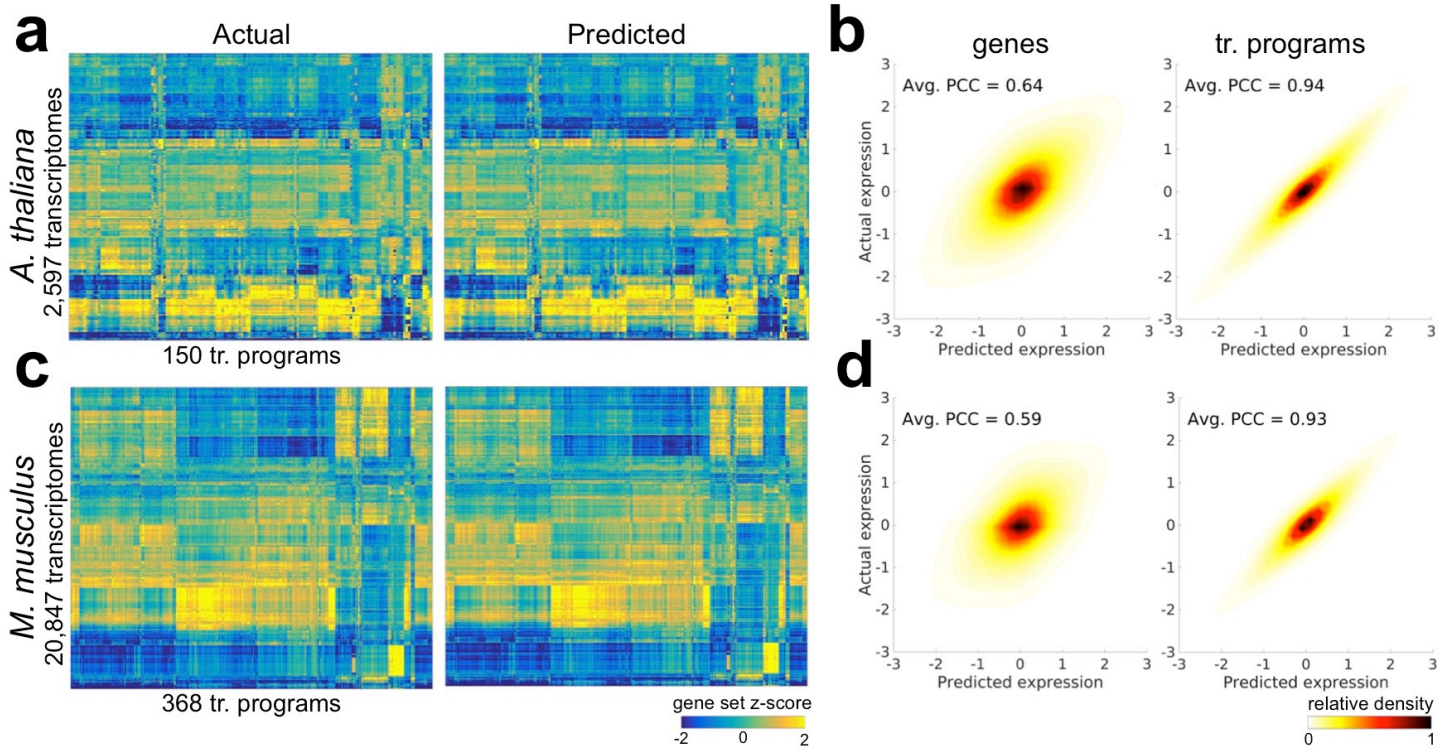

Figure 3. Tradict prospectively predicts unseen transcriptomes with robust accuracy. Tradict's prospective prediction accuracy during 20-fold cross validation of the entire training collection for both organisms. a) Heatmaps illustrating test-set reconstruction performance of all transcriptional programs for *A. thaliana*. b) Density plots of predicted vs. actual test-set expression for all genes (left) and transcriptional programs (right) for *A. thaliana*, after controlling for inter-sumbission biological signal. The intra-submission expression of each gene and transcriptional program was z-score transformed to make their expression comparable. c & d) Same as a & b, but for *M. musculus*.

(Supplemental Tables 3-4). We did not find any transcriptional programs with relatively low (PCC < 0.9) prediction accuracy that were not composed of a few or lowly expressed genes.

We additionally found Tradict's performance to be superior to several baselines. These include two approaches developed in Donner *et al.* (2012) for microarray[14], and a version of Tradict that uses the 100 most abundant genes as it's selected markers (Figure S5a, Supplemental Note 2). The latter baseline examines the utility of simple shallow sequencing. We additionally found Tradict's predictions were highly robust to noise (Figure S5b, Supplemental Note 2) -- a consequence of its probabilistic framework, in which training and prediction is performed in the space of denoised latent abundances.

Case studies reveal the power of predicting and studying predefined transcriptional programs

To demonstrate how Tradict may be applied in practice, we focused on two case studies related to innate immune signaling -- one performed using bulk *A. thaliana* seedlings (detailed below), and the other using primary immune *M. musculus* cell lines (detailed in Supplemental Note 3; Figure S6). We trained Tradict on our full collection of training transcriptomes for each organism to produce two organism-specific Tradict models. Each was based on the selection of 100 markers learned from the full training transcriptome collection (Supplemental Tables 5-6) that we assert are globally representative, and context-independent. The case study samples do not, of course, appear in the collection of training transcriptomes.

Tradict accurately predicts temporal transcriptomic expression patterns for a diverse panel of *A. thaliana* immune signaling mutants under different hormone perturbations - The hormones salicylic acid (SA) and jasmonic acid (JA) play a major, predominantly antagonistic regulatory role in the activation of plant defense responses to pathogens. Yang *et al.* (2016) investigated the effect of a transgenically expressed bacterial effector, HopBB1, on immune signaling in *A. thaliana*[28]. In their study, they performed a time course experiment, treating plants with MeJA (a JA response inducer), BTH (an SA mimic and SA response inducer), or mock buffer and monitored the transcriptome of bulk seedlings at 0 hr, 1 hr, 5 hr, and 8 hr post treatment. These experiments included several immune signaling mutants with differing degrees of response efficiency to MeJA and BTH treatment. Among other findings, they conclude that HopBB1 enhances the JA response, thereby

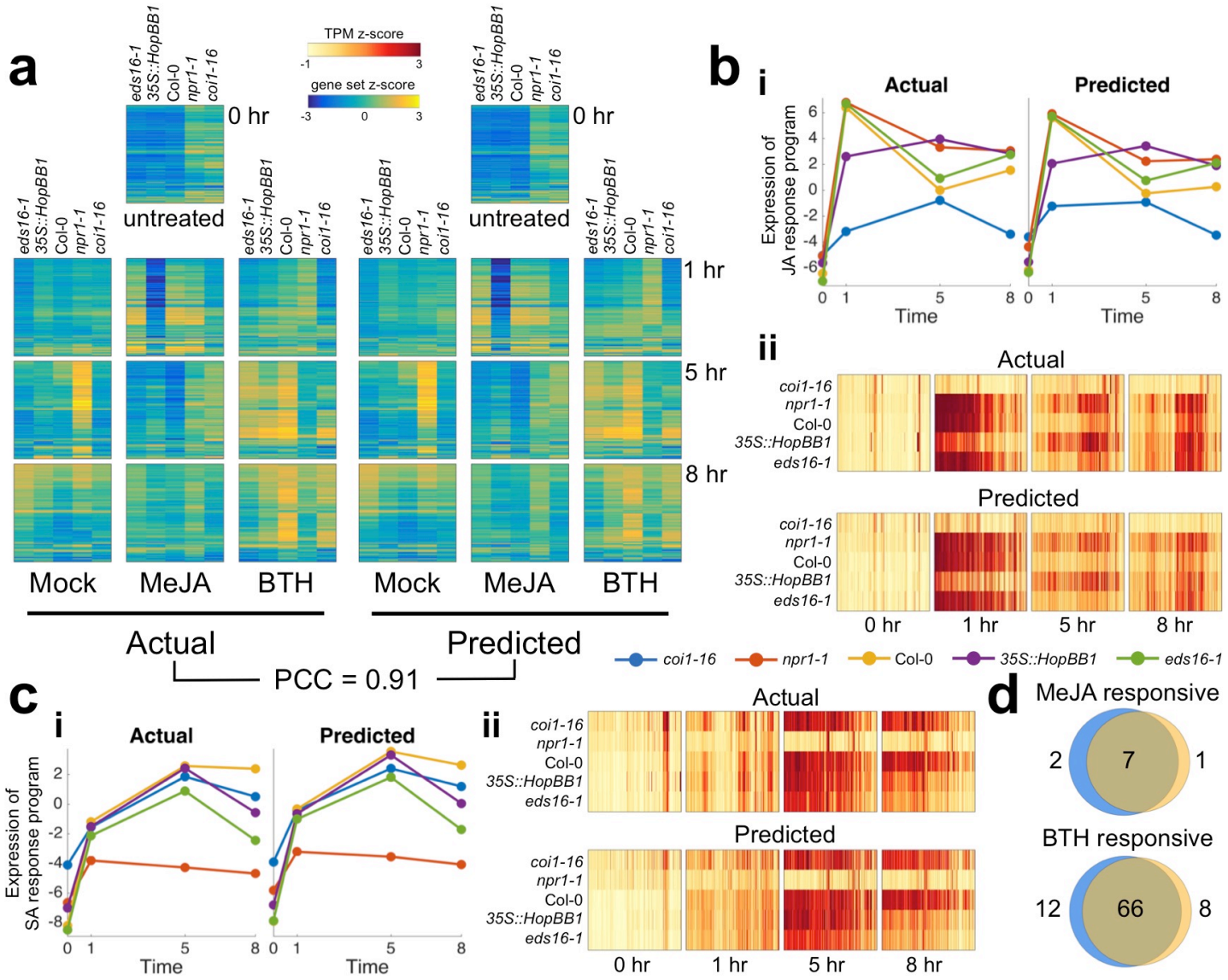

Figure 4. Tradict accurately predicts transcriptional responses across time in response to hormone perturbation in an *A. thaliana* innate immune signaling dataset. After being trained on the full *A. thaliana* training transcriptome collection, the selected set of 100 globally representative and context-independent markers were used to predict the expression of transcriptional programs and all genes for the transcriptomes presented in Yang *et al.* (2016). a) Actual vs. predicted heatmaps for the expression of all 150 transcriptional programs in *A. thaliana* across genotype, time, and hormone treatment. b) Predicted vs. actual expression of i) the JA response transcriptional program, and ii) the genes involved in the JA response program. c i-ii) Same as b, but for the SA response transcriptional program. d) Hypothesis free, differential transcriptional program expression analysis as performed on the actual expression of transcriptional programs vs those predicted by Tradict. Blue circles represent the actual and orange represent the predicted. All heatmaps are clustered in the same order across time, treatment, genotype, and between predicted and actual.

repressing the SA response and facilitating biotrophic pathogen infection.

We asked to what extent strategic undersampling of the transcriptome and application of Tradict could quantitatively recapitulate the findings of Yang *et al.* (2016). Given Tradict's near perfect accuracy on predicting the expression of transcriptional programs, we took a top down, but hypothesis driven approach to our analysis which first examined the expression of all transcriptional programs. Figure 4a illustrates the actual and predicted expression of all transcriptional programs in *A. thaliana* as a function of time and treatment. Here, Tradict reconstructs the expression of all transcriptional programs with an average PCC of 0.91.

Recall that the genes participating in each of our transcriptional programs are pre-defined, in this work, by a carefully chosen, interpretable, but maximally representative set of GO biological processes. Therefore, given the goals of this study, we next examined the expression of the "response to jasmonic acid" and "response to salicylic acid" transcriptional programs. Figure 4b shows the

expression behavior for the "response to jasmonic acid" transcriptional program across all the genotypes and time points upon MeJA treatment. More specifically, part (i) shows that the predicted expression and actual expression are qualitatively and quantitatively in agreement, both in magnitude and rank across the different genotypes. For example, as expected, *coi1-16*, which cannot sense JA, does not respond to the MeJA stimulus, while wildtype Col-0 does. However, even more subtle expression dynamics are captured by Tradict's predictions. For example, *eds16-1* and *npr1-1* -- slightly and strongly impaired SA responders, respectively -- are slightly and strongly hyper-responsive to MeJA, respectively -- just as expected from the JA-SA antagonism. Furthermore, as demonstrated in Yang *et al.* (2016), the *35S::HopBB1* transgenic line exhibits a prolonged and sustained JA response for both the actual and predicted expression for this transcriptional program. Part (ii) of Figure 4b illustrates the expression of all the MeJA responsive genes in this transcriptional program. Again Tradict's predictions are in good agreement with actuality, achieving a PCC of 0.72, and it's visually clear that the expression magnitude of these genes positively correlates with the registered expression magnitude of the "response to jasmonic acid" transcriptional program. Figure 4c parts (i) and (ii) are presented in the same light as Figure 4b, but are instead illustrated for the SA response transcriptional program and constituent genes under BTH treatment. Again predictions match actuality, and the observed trends make biological sense[29].

In order to illustrate Tradict's use in hypothesis-free investigation, we performed a differential transcriptional program expression analysis for transcriptional programs affected by MeJA or BTH treatment (Figure 4d, see Methods). Differentially expressed transcriptional programs based on Tradict's predictions versus actual measurements were highly concordant and biologically reasonable. Transcriptional programs differentially expressed with respect to MeJA treatment included "response to bacterium," "defense response to fungus", "response to wounding," and "response to jasmonic acid" as expected. Transcriptional programs differentially expressed with respect to BTH treatment included various abiotic stress responses, "defense response to fungus", "response to jasmonic acid" (via antagonism), and "response to salicylic acid," again, as expected.

Discussion

Tradict is an accurate, robust-to-noise algorithm for high fidelity transcriptome reconstruction given the expression measurements of a small, machine-learned subset of 100 marker genes. Given the comprehensiveness, stability, and exponentially growing size of the training datasets we have assembled from publicly available sources, the 100 markers Tradict learns are likely to be predictive independent of most contexts and applications. Furthermore, Tradict's ability to near perfectly model the expression of a biologically comprehensive, but interpretable list of annotated transcriptional programs enables one to rapidly generate hypotheses and dissect mechanism.

When coupled with target RNA sequencing, we believe Tradict can enable transcriptome-wide screening cheaply at scale. Commercial[19,20] and non-commercial[21,22] methods exist for targeted RNA sequencing in a multiplexed manner, and they are able to measure the expression of 10's-100's of genes with accuracy, making their use compatible with Tradict. Nevertheless, we see improvements in these approaches that could increase multiplexibility and reduce cost. More specifically, we estimate that Tradict coupled with a time and resource efficient targeted RNA-sequencing protocol could bring the cost of obtaining actionable transcriptome-wide information for thousands to tens of thousands of samples to close to $1 a sample (Supplemental Note 4).

This scale is exactly what is needed for many high-throughput profiling and screening applications. Many single-cell experiments now profile thousands of cells, and instead of using shallow sequencing (Figure S5), these investigations may benefit from a more targeted effort to query easier-to-measure, but statistically informative transcripts. Tradict could then use these measurements to generate predictions about the status of all transcriptional programs in the cell. With respect to screening, forward genetic screens in most eukyarotic organisms require assaying 10^3-10^4 mutants. Small molecule, or more generally chemogenomic, drug screens often require screening thousands of molecules against multiple cell lines in multiple conditions. Though in these cases the screen is made cheap and scalable by monitoring an easily selectable phenotype, new phenotyping architectures must be developed and optimized for each new screen (e.g. reporter lines, imaging hardware/software). Given the ubiquity of RNA, a transcriptome-wide screening approach does not suffer from such a drawback.

Furthermore, and more importantly, though quickly interpretable, the phenotype being screened for is usually a uni-dimensional datum that offers little

immediate insight into mechanism. In contrast, using Tradict to help perform transcriptome-wide screening could strongly couple the process of hypothesis generation and mechanistic investigation. Here, we argue that the scalable monitoring of the expression of a comprehensive list of just a few hundred transcriptional programs affords just the right balance of nuance and interpretability. Consequently, this efficient investigation, largely facilitated by Tradict, could greatly accelerate the pace of genetic dissection, breeding, and drug discovery.

Acknowledgements
We thank Brian Cleary, Aviv Regev, Amaro Taylor-Weiner, Craig Bohrson, and Derek Lundberg for valuable discussions in preparing this manuscript. SB was supported by a Churchill Scholarship from the Winston Churchill Foundation of the United States and by a training grant from the NHGRI/NIH (T32 HG002295). PJPLT was supported by a fellowship from the Pew Latin American Fellows Program in the Biomedical Sciences. This work was additionally funded, in part, by a fellowship to PW from the Gatsby Foundation (GAT2373/GLB) and by grants to JLD from the National Institutes of Health (1RO1 GM107444), the Gordon and Betty Moore Foundation (GBMF3030), and the HHMI. JLD is an Investigator of the Howard Hughes Medical Institute.

Code availability
A MATLAB implementation of Tradict is available at https://github.com/surgebiswas/tradict. All code to perform data downloads, analysis, and generate figures are available at https://github.com/surgebiswas/transcriptome_compression. Note that in order to make Tradict open-source and reach a wider audience we are developing user-friendly, unit tested R-package of Tradict that will be made available before publication.

References

1. Jacob, F. & Monod, J. Genetic regulatory mechanisms in the synthesis of proteins. *J. Mol. Biol.* **3,** 318–356 (1961).
2. Kaufmann, K., Pajoro, A. & Angenent, G. C. Regulation of transcription in plants: mechanisms controlling developmental switches. *Nat. Rev. Genet.* **11,** 830–842 (2010).
3. Mitchell, P. J. & Tjian, R. Transcriptional Regulation in Mammalian Cells by DNA Binding Proteins. *Science (80-.).* **245,** 371–378 (1989).
4. Segal, E. *et al.* Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat. Genet.* **34,** (2003).
5. Hart, Y. *et al.* Inferring biological tasks using Pareto analysis of high-dimensional data. *Nat. Methods* **12,** (2015).
6. Shoval, O. *et al.* Evolutionary Trade-Offs, Pareto Optimality, and the Geometry of Phenotype Space. *Science (80-.).* **336,** 1157–1160 (2012).
7. Jeong, H., Tombor, B., Albert, R., Oltvai, Z. N. & Barabási, a L. The large-scale organization of metabolic networks. *Nature* **407,** 651–654 (2000).
8. Albert, R., Lee, J. H. & Barabási, A.-L. Error and attack tolerance of complex networks. *Nature* **406,** 378–382 (2000).
9. Barabási, A.-L. & Albert, R. Emergence of Scaling in Random Networks. *Science (80-.).* **286,** 509–513 (1999).
10. Troyanskaya, O. *et al.* Missing value estimation methods for DNA microarrays. **17,** 520–525 (2001).
11. Liew, A. W., Law, N. & Yan, H. Missing value imputation for gene expression data : computational techniques to recover missing data from available information. **12,** 498–513 (2010).
12. Celton, M., Malpertuy, A., Lelandais, G. & Brevern, A. G. De. Comparative analysis of missing value imputation methods to improve clustering and interpretation of microarray experiments. (2010).
13. Ling, M. H. T. & Poh, C. L. A predictor for predicting Escherichia coli transcriptome and the effects of gene perturbations. *BMC Bioinformatics* **15,** 140 (2014).
14. Donner, Y., Feng, T., Benoist, C. & Koller, D. Imputing gene expression from selectively reduced probe sets. *Nat. Methods* **9,** (2012).
15. Heimberg, G. *et al.* Low Dimensionality in Gene Expression Data Enables the Accurate Extraction of Transcriptional Programs from Shallow Sequencing Article Low Dimensionality in Gene Expression Data Enables the Accurate Extraction of Transcriptional Programs from Shallow Sequencing. *Cell Syst.* **2,** 239–250 (2016).
16. Pollen, A. A. *et al.* Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. *Nat. Biotechnol.* **32,** (2014).
17. Kliebenstein, D. J. Exploring the shallow end; estimating information content in transcriptomics studies. *Front. Plant Sci.* **3,** 1–10 (2012).

18. Jaitin, D. A. *et al.* Massively Parallel Single-Cell RNA-Seq for Marker-Free Decomposition of Tissues into Cell Types. *Science (80-.).* **343,** 776–779 (2014).
19. ThermoFisher Scientific. Targeted RNA Sequencing by Ion Torrent Next-Generation Sequencing. at <http://www.thermofisher.com/us/en/home/life-science/sequencing/rna-sequencing/targeted-rna-sequencing-ion-torrent-next-generation-sequencing.html>
20. Illumina. TruSeq Targeted RNA Expression Kits. at <http://www.illumina.com/products/truseq-targeted-rna-expression-kits.html>
21. Mercer, T. R. *et al.* Targeted sequencing for gene discovery and quantification using RNA CaptureSeq. *Nat. Protoc.* **9,** 989–1009 (2014).
22. Fu, G. K. *et al.* Molecular indexing enables quantitative targeted RNA sequencing and reveals poor efficiencies in standard library preparations. **111,** 1891–1896 (2014).
23. Biswas, S. The latent logarithm. *arXiv* 1–11 (2016).
24. Ma, S. & Kosorok, M. R. Identification of differential gene pathways with principal component analysis. *Bioinformatics* **25,** 882–889 (2009).
25. Fan, J. *et al.* Characterizing transcriptional heterogeneity through pathway and gene set overdispersion analysis. *Nat. Methods* **13,** 241–244 (2016).
26. Tropp, J. a & Gilbert, A. C. Signal Recovery From Random Measurements Via Orthogonal Matching Pursuit. *IEEE Trans. Inf. Theory* **53,** 4655–4666 (2007).
27. Tropp, J. a., Gilbert, A. C. & Strauss, M. J. Algorithms for simultaneous sparse approximation. Part I: Greedy pursuit. *Signal Processing* **86,** 572–588 (2006).
28. Yang, L. *et al.* The Pseudomonas syringae type III effector HopBB1 fine tunes pathogen virulence by gluing together host transcriptional regulators for degradation. *Submitted* (2016).
29. Jones, J. D. G. & Dangl, J. L. The plant immune system. *Nature* **444,** 323–9 (2006).
30. New England BioLabs Inc. SplintR Ligase. at <https://www.neb.com/products/m0375-splintr-ligase>
31. Lohman, G. J. S., Zhang, Y., Zhelkovsky, A. M., Cantor, E. J. & Jr, T. C. E. Efficient DNA ligation in DNA – RNA hybrid helices by Chlorella virus DNA ligase. *Nucleic Acids Res.* 1–14 (2013). doi:10.1093/nar/gkt1032
32. Lundberg, D. S., Yourstone, S., Mieczkowski, P., Jones, C. D. & Dangl, J. L. Practical innovations for high-throughput amplicon sequencing. *Nat. Methods* **10,** 999–1002 (2013).